

## Examination of the Validity of Auditory Traits and Tests

Gregory A. Flamme, PhD

### Introduction

During recent years, the field of rehabilitative audiology has been moving toward the development and systematic use of questionnaires and psychophysical tests as indicators of clinical and research outcomes. As a result, these instruments now occupy a greater role in clinical decisions, either directly through application in the clinic or indirectly through research projects aimed at selecting acceptable and/or evidence-based clinical activities. As greater importance is awarded to the results of studies using questionnaire and psychophysical outcome measures, clinicians and clinical researchers must take more responsibility for the informed selection of the measures.

Gatehouse (2001) notes in this issue of *Trends in Amplification* that third-party funding agencies typically require evidence that a service was beneficial as a condition for reimbursement. He further recommends that practitioners monitor their clinical procedures through the evaluation of quantitative data and empirical evidence of the benefit of a service. This evidence will likely

be sought by examining whether a given service is likely to provide a better outcome for the listener than some less expensive service or no service at all. To this end, the characteristics of the test(s) used to examine the outcome must be known. Does the test score actually represent performance in the domain, or area, that the tester wishes to explore? Is this domain distinctly different from other domains that the tester plans to also explore? What external factors are likely to influence the scores from the test? What is the likelihood that a given test will provide a meaningful and different type of information about the listener's hearing status?

The answers to these questions inform the tester about the validity of the domain, or the *trait*, that the tester wishes to know about, and the specific test that the tester has chosen to provide an estimate of that trait. Test reliability, validity, and the relationships between the quantities that the investigator wishes to estimate (i.e., the object of the measurement) are not trivial, and the careful evaluation of tests involves research designs that are not commonly used in audiology.

---

University of Memphis, Memphis, Tennessee

Financial support for this project was provided by the Center for Research Initiatives and Strategies for the Communicatively Impaired (CRISCI) at the University of Memphis School of Audiology and Speech-Language Pathology

Correspondence: Greg Flamme, PhD, Department of Speech Pathology and Audiology, University of Iowa, 250 Hawkins Drive, Iowa City, IA 52242

©2001 Westminster Publications, Inc., 708 Glen Cove Avenue, Glen Head, NY 11545, U.S.A.

There is relatively little information about the quantitative characteristics of the scores from the array of tests conventionally used in clinical audiology and auditory research. It is possible that many of the tests that are currently used in audiology measure the same aspect of hearing, but the scores returned by these tests may be different because they are influenced by different factors that are unrelated to the person's underlying ability to hear. Such tests provide the same information about the person's hearing status, but differences in subject responses to the measurement methods used in each test could lead the clinician or clinical researcher to incorrectly conclude that fundamental hearing-related differences exist between the listeners.

A distinction must be made between the *measurement operation* that returns an estimate of a trait and the trait itself. Other things also influence the outcome of a test, including random error and the *measurement method*. Only a few of these influences are of interest to the tester, and the remaining influences constitute systematic bias and random error in the estimate. A *test* is a measurement operation encompassing stimuli, presentation levels, listener instructions, response criteria, scoring algorithms, question formats, etc. Only a few aspects of a measurement operation are relevant to the trait, while others are elements that must be controlled to obtain reliable data, but are not influenced by the trait.

In audiology and auditory science, many traits have been identified (e.g., hearing sensitivity, directional hearing, speech understanding), and many measurement operations have been developed to estimate a listener's performance on these traits. Given the large number of auditory traits, clinicians and clinical researchers are often left wondering which trait(s) to include in an evaluation protocol, and which tests will give the best estimate of these traits. After making the decision about the traits one wishes to estimate, another decision must be made about which test (i.e., measurement operation) to use as an estimate of each trait. For example, if an estimate of speech understanding in noise is desired, the test protocol developer can select from the Connected Speech Test (CST), (Cox *et al.*, 1987a, 1987b; 1988; 1989), the Hearing-in-Noise test (HINT), (Nilsson *et al.*, 1994), the Speech Perception in Noise (SPIN) test (Kalikow *et al.*, 1977) along with many others. Note that these tests use *psychophysical*

*methods*, wherein an acoustic stimulus is presented and the listener is asked to report the word(s) that he or she was able to hear. The psychophysical method is only one of the general approaches to estimating a listener's performance on a trait. Estimates of a trait are also commonly obtained through *self-report methods*. Self report methods involve a listener's estimate of hearing performance in a specified situation. Self-report tests of speech understanding in noise include the background noise (BN) subscale of the Profile of Hearing Aid Benefit (Cox *et al.*, 1991) and many others (see Bentler and Kramer, 2000, for a comprehensive list).

Another approach to estimating listener performance on a trait is through *significant-other report methods*. With this method, people likely to be familiar with a listener's performance are asked to estimate the listener's hearing performance in a specified situation. This method has been used in the Significant-Other Assessment of Communication (SOAC), (Schow and Nerbonne, 1982), and the Nursing Home Hearing Handicap Index (Schow and Nerbonne, 1977), among others.

The existence of a large number of tests for estimating a given trait provides both an advantage and a disadvantage. The advantage is that the tester may select a test with measurement operations that are well suited to the research protocol or the clinical environment. The disadvantage is that two tests intended to estimate performance on the same trait might not actually estimate the same thing. Information about the validity of each test of a trait is necessary to evaluate this potential disadvantage.

The validity evidence typically found in the literature can be described as *convergent validity*. Convergent validity data show the amount of shared variance between tests intended to estimate performance on the same trait. Criterion validity studies provide convergent validity data. In this type of study, the correlation between two tests presumed to estimate the same trait is used to estimate the amount of shared variance between the tests. The estimate of shared variance from a criterion validity study includes the shared variance between the traits, that is, *trait variance*, confounded with the shared variance between the measurement methods (i.e., the systematic data patterns arising from the similarities in measurement operations, irrespective of the trait that the tester wishes to estimate, which is called *method variance*).

To separate the *trait variance component* of a given test score from the *method variance component*, it is necessary to administer at least two additional tests, one intended to estimate the same trait but using a different measurement method, and another estimating a different trait but using a similar measurement method. The covariance between tests intended to estimate the same trait provides an estimate of the trait variance in the test. The covariance between tests using the same method provides an estimate of the amount of method variance in the test. This approach to partitioning trait and method variance is the foundation of what is known as a *multitrait-multimethod* (MTMM) research design (Campbell and Fiske, 1959; Nunnally and Bernstein, 1994; Widaman, 1992), which was used in this study.

After the trait and method components of a test score have been separated, it is possible to obtain an unbiased estimate of the amount of shared variance across traits. This type of information is called *discriminant validity*. Discriminant validity was defined by Nunnally and Bernstein (1994) as the ability of a test of a trait to "produce relevant group differences." Discriminant validity information is important because even if a pair of measurement operations returned perfect estimates of their respective traits, one cannot justify the inclusion of both tests in a protocol if the traits are perfectly correlated. In a more realistic example, where each test returns an imperfect estimate of the trait, a tester is liable to interpret between-test differences in terms of the trait, although the actual cause of the score differences might have been an unrelated factor, such as a change in the listener's response criterion (Cronbach and Meehl, 1955). Discriminant validity data are rarely reported in the auditory literature.

*Construct validity* studies integrate evidence of convergent and discriminant validity into an evaluation of whether there is empirical evidence for a construct, whether constructs can be distinguished from one another, and an evaluation of whether a given test returns an estimate of its intended trait. As such, construct validation is a central issue in the exploration of an area of knowledge and in test development. A construct is a dimension upon which people are expected to differ. For example, a substantive trait (e.g., speech understanding in noise) is a type of construct. However, individual differences in response to a measurement method also constitute

a dimension of intersubject differences, and therefore differences related to the measurement method also can be considered a construct.

Construct validation procedures provide many types of information. First, they provide an unbiased estimate of the relatedness of traits. Second, they provide an unbiased estimate of the relatedness of methods. Third, they provide an estimate of the size of trait and method variance components in scores from a test intended to estimate a listener's ability on a given trait. Knowledge of the relatedness of different constructs is helpful to clinicians and clinical researchers because a data collection protocol, whether for clinical or research purposes, is most efficient when weak relationships exist among the estimated traits. As the traits estimated by a set of tests become more closely related, the meaningful aspects of the test scores become redundant, and only add to the time it takes to complete the test protocol. Conversely, if the traits are unrelated, each test provides unique information about the state of the listener, providing a richer representation of the listener's characteristics. Knowledge about the relationships between traits also advances theory in the topic area through the *nomological net* (Cronbach and Meehl, 1955), which is an orderly exploration of the relationships among traits within a given area of knowledge. Estimation of the relationships between the influences of different measurement methods is important because the integration of results across tests using strongly related measurement methods will have a systematic influence on the results of a study.

Information about the relative sizes of the trait and method variance components of a test can help a tester select the most appropriate test from an array of available tests. Tests returning scores having relatively large trait variance components will obtain better estimates of a listener's ability on the trait than tests with smaller trait components. Tests with relatively large method variance components will provide a relatively good estimate of the listener's response to the measurement operation; however, they will provide relatively little information about the *object of measurement* (i.e., performance on the trait). Furthermore, performance differences on the trait will be difficult to detect using a test that is greatly influenced by measurement method. Finally, tests having both small trait and small method variance components will either contain a great

deal of unique variance (i.e., variance that is systematic, but cannot be attributed to the intended trait), or they will contain little systematic variance, and thus the test will return scores that are random and meaningless.

The relationships between constructs and the relative amounts of these variance components in common auditory tests can inform the development of clinical and research protocols. However, a review of the audiology literature found no studies separating trait, method, and unique/error components of test scores, or studies providing an unbiased estimate of the relationships among auditory traits or among common methods used in auditory tests. The current study provides these data for a set of three traits (direction and distance hearing, soft sounds hearing, speech understanding in noise), and a set of three methods (psychophysical, self-report, and significant-other report).

---

#### Direction and Distance Hearing

---

The trait of *direction and distance* (DD) hearing represents a listener's ability to determine the location of a sound source, in terms of azimuth, elevation, and distance. Many factors appear to influence performance on this trait, including interaural time differences, interaural level differences, a listener's awareness of the acoustic environment, and signal audibility (Blauert, 1997; Wightman and Kistler, 1997).

DD hearing can be estimated via many measurement operations, including psychophysical approaches, self-report questionnaires, and significant-other report questionnaires. Although much research has focused on DD hearing in quiet environments, a few studies have estimated listener performance in noise, which is of considerable interest because typical listening environments contain noise (Pearsons *et al.*, 1977). Two studies (Good and Gilkey, 1996; Lorenzi *et al.*, 1999a) provide systematic evaluations of DD hearing in normal hearers across a range of signal-to-noise ratios (SNR), using a click train target stimulus and a white noise masker. The results of both studies indicated that DD hearing was resistant to the effects of noise; with nearly perfect performance obtained at SNR of 0 to 2 dB and greater. In addition, both studies observed considerable amounts of, and different types of, response bias across subjects.

Lorenzi and associates (1999b) also estimated DD hearing in noise in listeners with sensorineur-

al hearing impairments, using the same procedures as Lorenzi and associates (1999a). The results were similar to the results obtained with normal hearers; DD hearing was not influenced by the masker until the signal was below about 0 dB SNR. However, in some listeners, DD hearing performance never reached perfect accuracy, even in quiet. These listeners tended to have poorer pure tone thresholds, but further data collection and analyses revealed that the performance decrement involved more than stimulus detectability.

Self-report methods have also been used to estimate DD hearing ability (Flamme *et al.*, 1999; Kramer *et al.*, 1995; Noble and Atherley, 1970; Noble *et al.*, 1995). There are a few commonalities among the questionnaires used in these studies. First, although the instruments have different names, many of them have very similar items. Second, in the studies where the reliability of the questionnaire has been evaluated (Flamme *et al.*, 1999; Kramer *et al.*, 1995; Noble and Atherley, 1970), the scores from the DD hearing scales tend to have high reliability.

Some self-report questionnaire measures of DD hearing have been correlated with psychophysically-measured DD hearing. Noble and Atherley (1970) and Kramer and associates (1995) both observed moderate correlations between their respective self report estimates of DD hearing and psychophysically measured DD hearing. However, moderate correlations were also noted between self-report estimates of DD hearing and various psychophysical measures of the ability to hear soft sounds (i.e., pure tone thresholds) and understand speech in quiet (i.e., speech reception thresholds). Assuming that the traits of DD hearing and soft sounds (SS) hearing are not strongly related, one would expect measures of the same trait to correlate more strongly with one another than measures of different traits. It is unknown whether the lack of a unique relationship between measures of DD hearing is due to strong underlying relationships between the traits of DD hearing, SS hearing, and speech understanding in quiet, or if the observed correlations indicate a problem with one or both of the measures of DD hearing.

---

#### Soft Sounds Hearing

---

*Soft sounds* (SS) hearing represents a listener's ability to detect low-level signals in the environment. SS hearing has typically been included in clinical test protocols, both through unstructured

interview questions and through psychophysical measures of pure tone thresholds. Estimates of SS hearing are often considered a poor indicator of the communication problems experienced by listeners in typical environments (Erdman, 1994; Erdman and Demorest, 1998; Swan and Gatehouse, 1990). Moderate correlations are regularly obtained between psychophysically measured estimates of SS hearing (i.e., pure tone thresholds) and hearing disability and handicap. This is often interpreted as an indication that pure tone thresholds fail to include important factors that impact function in daily life, for example, the noise environment, listener compensatory strategies, and so on (Erdman and Demorest, 1998).

Although SS hearing is typically estimated using psychophysical methods, questionnaires can also serve as indicators of SS hearing. Coren and Hakstian (1992) report the development of the Hearing Screening Inventory (HSI), a 12-item self-report questionnaire designed to correlate strongly with psychophysically measured estimates of SS hearing. Coren and Hakstian (1992) reported a strong correlation ( $r = .81$ ) between the HSI scale score and bilateral four-frequency average hearing level (4FAHL), at mean thresholds across 500, 1000, 2000, and 4000 Hz.

### Speech Understanding in Noise

The third trait examined in this study is speech understanding in noise, or *understanding in noise* (UN) hearing. Many tests have been developed to estimate UN hearing performance, and the tests encompass psychophysical, self-report, and significant-other report methods. Cox and associates developed the Connected Speech Test (CST), which uses a psychophysical method. This test uses conversationally produced speech by a talker who demonstrated median intelligibility across a number of listening environments (Cox and associates, 1987a), and the test contains a large number of equivalent forms and sufficient sensitivity to detect a SNR change of 2 dB. Normative data were obtained for the CST test with people having normal hearing (Cox *et al.*, 1987b) and hearing impairments (Cox *et al.*, 1988). The internal consistency reliability of the CST is high ( $\alpha = .98$ ).

Certain subscales of the Profile of Hearing Aid Benefit (PHAB) (Cox *et al.*, 1991) are likely to provide a good estimate of UN hearing ability

using self-report methods. The PHAB is a modified form of the Profile of Hearing Aid Performance (PHAP) (Cox and Gilmore, 1990). The PHAP questionnaire was designed to estimate performance on a number of traits with a number of hearing aid wearers. The PHAB includes listener judgments of performance with aided and unaided listeners; the difference between these scores provides an estimate of hearing aid benefit. The PHAP/PHAB contains multiple subscales, estimating performance with familiar talkers, ease of communication, understanding speech in background noise, hearing under conditions of reduced cues, hearing in reverberant conditions, the level of distortion of sounds, and the aversiveness of sounds. An evaluation of the internal consistency reliability of these subscales indicates a reasonable level of internal consistency reliability ( $\alpha = .85-.91$ , across subscales). As estimates of UN hearing, only the background noise and reverberation subscales are relevant to the current study.

### Measurement Methods

Psychophysical methods have conventionally been regarded as gold standard or objective approaches to estimating auditory traits. But psychophysical methods are as susceptible to influences from irrelevant factors as other estimation methods. Choices regarding stimuli, listening environments, response criteria, and experiment duration might be irrelevant to the trait of interest, but may create systematic score differences among subjects. Psychophysical methods have the advantage of known stimuli, but they have the disadvantage that the stimuli are not necessarily representative of the stimuli presented to a listener in his/her daily life.

Self-report methods have the practical advantage of requiring little instrumentation (i.e., the test form and a pencil), and they provide greater assurance that the responses represent the listener's daily life experience. However, self-report methods lack well-defined stimuli, and although listening conditions may be reliably categorized across people, one cannot be certain that these listening conditions are acoustically similar.

Significant others play a major role in encouraging people with hearing impairment to seek help, possibly because hearing impairments cause a participation restriction for the hearing-impaired person's conversation partners. O'Mahoney

and co-workers (1996) found that only about 25% of people consulting a professional about a hearing problem were self-motivated. The majority of their subjects were motivated by family members or friends to address their hearing difficulties. Because significant others play such a large role in a person's choice to seek help for their hearing problem, the accuracy with which they judge the performance of a listener should be evaluated. A number of investigators (Chmiel and Jerger, 1993; Lormore and Stephens, 1994; Newman and Weinstein, 1986; Schow and Nerbonne, 1977, 1982; Stephens *et al.*, 1995) have investigated the abilities of significant others to judge the communication performance of people with hearing problems. In general, research involving the reports of significant others has used modified tests that were originally developed as self-report tests. In most cases, the modification consisted of the replacement of first-person statements (e.g., "I") with third-person (e.g., "he" or "she") statements. With the exception of the SOAC (Schow and Nerbonne, 1982), no psychometric evaluations of the modified instruments were reported. Schow and Nerbonne evaluated the test-retest reliability of the SOAC and found it to be acceptable ( $r = .90$ ).

The current study was designed to examine the construct validity of three traits (DD hearing, SS hearing, and UN hearing), and three measurement methods (psychophysical, self-report, and significant-other report). A multitrait-multimethod design encompassing three traits and three methods requires the administration of nine measures, one for each combination of trait and method. Although measures of each trait using psychophysical and self-report methods were readily available, measures using the significant-other report method were developed for use in this study. This development process followed the conventional approach of replacing first-person statements with third-person statements.

## Method

### General Procedure

Hearing-impaired adults and significant others participated in this study. For each participant three estimates of listener performance on each trait were obtained. Each test of a given trait used

a different method (psychophysical, self-report, and significant-other report). Thus, each hearing-impaired participant completed three psychophysical and three self-report questionnaire measures. Significant others (SO) completed three questionnaire measures. Self-report questionnaire measures were always completed before psychophysical measures were conducted. All questionnaires were completed in a paper and pencil format. To counteract potential order effects, measurement order was randomized across subjects, within each measurement method.

Psychophysical tests took place in a  $1.5 \times 2.2$  m single-walled sound booth with ambient sound levels sufficient for insert earphone threshold testing below 0 dB HL at frequencies above 250 Hz (ANSI S3.1, 1998). Reverberation times ( $RT_{60}$ ) measured at the listener's location in the sound booth were 129, 46, 37, 43, 49, and 51 msec for 1/3 octave band signals surrounding 250, 500, 1000, 2000, 4000, and 8000 Hz, respectively.

### Participants

Because this study evaluated patterns of inter-subject differences, few inclusion or exclusion criteria were used. A narrowly defined subject population might have resulted in a restricted response range, potentially biasing relationships. To participate in the study, subjects with hearing impairments needed to report an ability to judge their performance without hearing aids, regardless of whether they typically wore hearing aids. To participate in the study, significant others had to classify their amount of hearing difficulty as none or mild, on a scale of none, mild, moderate, moderately severe, and severe. No gender, age, or education criteria were considered for inclusion. All participants were paid.

Because the questionnaires are written in English, only individuals who used English as their primary language participated. The Flesh-Kincaid reading level of the questionnaire items was measured using Microsoft Word 97. The questionnaire with the highest grade level was the Profile of Hearing Problems (grade 8.1), which is described below. To ensure ability to read and understand questionnaire items, all participants demonstrated an ability to read above the ninth grade failure reading level on the Woodcock Word Identification test (Woodcock, 1973).

In the factor analysis literature, there is a lack of consensus about the sample size necessary to obtain a stable and accurate estimate of population values. Nunnally (1978) and Cattell (1978) offered some rules of thumb; however, these rules of thumb find little support in empirical evaluations (Arrindell and van der Ende, 1985; Barrett and Kline, 1981; Guadagnoli and Velicer, 1988). The accuracy and stability of a factor analytic solution involves more than sample size (MacCallum *et al.*, 1996; MacCallum *et al.*, 1999). Communalities also play an important role (Guadagnoli and Velicer, 1988; MacCallum *et al.*, 1999). Communalities represent the proportion of variance in a variable that is associated with common factors. High communalities (e.g., values of .80 or greater, indicating that 80% of the variance in a variable is associated with the common factors) indicate that the variable is highly influenced by its causal factor(s). A smaller sample is required when variables have high communalities. MacCallum and associates (1999) found that sample size has a small impact on the precision of parameter estimates when communalities are greater than about .60.

Communality values vary with the substantive area of research and the data set under analysis. For this reason, *a priori* estimates of the required sample size could not be made. The power for detecting a single correlation was used to determine this study's sample size. A target sample size of 50 was selected because an absolute correlation magnitude of  $r = .30$  was arbitrarily decided to be the smallest interesting effect. A correlation of  $r = .30$  in the MTMM correlation matrix would suggest that only 9% of the variance in one test can be predicted based on another measure. Given a Type I error level of .05, and a population correlation magnitude of .30, a sample size of 50 provided a power level between .5 and .6, with greater power to detect larger effects. For example, if the effect size were  $r = .40$ , the estimated power would be between .80 and .90 (Rosenthal and Rosnow, 1991).

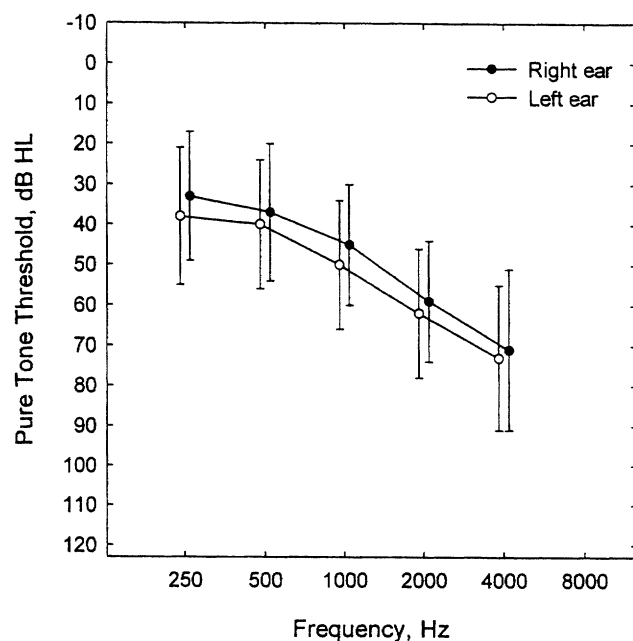
### Hearing-Impaired

Fifty-two of the 170 people asked to participate in this study volunteered. Two subjects dropped out due to health problems, another subject's significant other failed to meet the hearing difficulty inclusion

criterion, leaving 49 subjects with hearing impairments and their significant others in the sample.

Hearing aid ownership was not an inclusion or exclusion criterion. The majority of hearing-impaired listeners (74%) judged their unaided hearing difficulty to be either moderate or moderately severe. Approximately equal numbers of hearing-impaired subjects judged their unaided hearing difficulty to be either mild or severe, with 14% and 12% of subjects, respectively. The average age of hearing-impaired participants in this study was 73.9 (SD: 11.2). Approximately 29% of the hearing-impaired subjects in this sample were female.

The subjects in this study had sensorineural hearing loss. The majority (92%) of the hearing-impaired participants in this study had symmetric hearing loss, as defined by between-ear 4FAHL (mean pure tone threshold across 500, 1000, 2000, and 4000 Hz) differences less than or equal to 15 dB. Figure 1 represents the average pure tone thresholds, across hearing-impaired participants in this study. Standard deviations are also represented. Additional pure tone threshold information is provided in the Results section.



**Figure 1.** Mean pure tone thresholds for participants with hearing impairment (error bars represent  $\pm 1$  SD).

### Significant Others

Significant others fit into three basic categories, spouses ( $n = 37$ ), children/relatives ( $n = 7$ ), and close friends ( $n = 5$ ). The mean age of spouses was 68.0 (SD: 12.4). The mean age of children/relatives was 47.9 (SD: 11.9). The mean age of close friends was 71.5 (SD: 6.2). Significant others were predominately female (87%, 86%, and 100% for spouse, child/relative, and close friend subgroups, respectively).

### Tests

All sound field stimuli were measured at the location in the test room that was occupied by the center of the listener's head, hereafter referred to as the "reference position." Stimuli delivered via earphones were calibrated in a HA-1 2cc coupler. Each psychophysical stimulus was calibrated at the beginning of each test day.

### Direction and Distance Hearing

Psychophysical test stimuli. The target signal was a train of 23- $\mu$ sec impulses, repeated at a rate of 100 Hz and digitally low-pass filtered at 11 kHz. The click train duration was approximately 300 msec. The masker was white noise, digitally low-pass filtered at 14 kHz. The noise masker duration was approximately 900 msec, and was gated on and off using a 25-msec Hanning function. Signals were filtered to produce a flat 1/3-octave band frequency response between 100 and 16,000 Hz. The root mean square (RMS) deviations to a flat spectrum were less than 1 dB across this frequency range. One-third octave band signal and noise spectra were matched within 0.5 dB between 125 and 16,000 Hz.

The click train was presented at an overall level of 64 dBA at the reference position; the masker was presented at 66 dBA. The level of the click train was selected to match the overall level of the target talker in the psychophysical test of UN hearing. The signal-to-noise ratio of 2 dB was selected to ensure intersubject score differences, which are necessary in this type of research. To eliminate the possible use of low-level system noises for localization cues, a white noise signal was equalized to produce a flat ( $\pm 2$  dB) spectrum between 125 and 10,000 Hz, amplified, and rout-

ed to a loudspeaker located directly above the loudspeaker at 0-degree azimuth in the sound booth. This continuous low level signal was presented at 54 dBA at the reference position. At each 1/3 octave band between 125 and 10,000 Hz, this white noise was at least 15 dB above any system noise.

Psychophysical test apparatus. The click train was converted into analog form using a 32-bit sound card on board a 266-MHz PC computer. The 900-msec white noise was generated in MATLAB as a vector of normally distributed random numbers. The continuous low-level white noise masker was generated by Tucker-Davis Technologies (TDT) WG1 waveform generator. The 900-msec noise was gated using the MATLAB hanning.m software function. Because short-duration white noise samples were used, any single sample could deviate from a flat spectrum. To control for these deviations, ten tokens of white noise were generated. Ten stereo .wav sound files were created within MATLAB. The same click train waveform was used in one channel of each file; different noise tokens were stored in the other channel in each file.

TDT PA4 programmable attenuators, Klark-Teknik DN360 graphic equalizers, Crown D-150 power amplifiers, and Optimus Pro 7 AV loudspeakers were used to obtain the desired overall signal levels at the reference position. Seven loudspeakers were placed in 30-degree increments from -90 degrees azimuth to +90 degrees azimuth, with the loudspeaker faces 1 m from the reference position, at 0 degrees elevation.

Psychophysical test procedure. The psychophysical DD hearing test consisted of 14 practice stimulus presentations (i.e., two per loudspeaker) and 140 experimental stimulus presentations (i.e., 20 per loudspeaker). For each presentation, the amplified click train was randomly routed to one of seven loudspeakers. The 900-msec white noise was routed to one of the loudspeakers, at either 90 degrees or -90 degrees azimuths. Hence, both signals came from the same loudspeaker in one seventh of the trials. The practice run procedure was the same as the experimental run, but with a +2 dB SNR. Subjects were asked to report the loudspeaker from which the buzzing sound came; responses were given via a keypad, and automatically entered into a computerized scoring program. No feedback was given during the practice or experimental stimulus presentations, but in a few cases, a separate

practice run without the 900-msec noise was provided to help the listener discriminate the two signals. The low-level white noise masker was presented continuously throughout the experimental run.

**Questionnaires.** Self-report indications of DD hearing were obtained using the 14 item LOCATE questionnaire (Flamme *et al.*, 1999). Instructions asked the participants to answer each question as if they were not wearing hearing aids, if they owned any. Significant-other report estimates of DD hearing were obtained using the LOCATE-SO questionnaire. This questionnaire was identical to the standard LOCATE form, except that "I" statements were replaced by "he/she" statements. Instructions stated that participants were to answer each question as if the hearing-impaired person were not wearing hearing aids, if the hearing-impaired person owned any. Instructions and questionnaire forms are reproduced in Flamme (2000).

### Soft Sounds Hearing

**Psychophysical test stimuli.** Pure tones were presented at audiometric test frequencies between 250 and 8000 Hz. Tones were manually pulsed with a minimum 200-msec duty cycle (i.e., 200/200). Intensity was adjusted in 5-dB increments.

**Psychophysical test apparatus.** Listeners were seated in a single-walled sound booth during measurements. Stimuli were generated with a Madsen OB-922 clinical audiometer and routed through ER-3A insert earphones.

**Psychophysical test procedure.** Threshold for a given stimulus was obtained using standard clinical procedures. Threshold was defined as the lowest level audible to the subject on at least two of three trials (ANSI, 1986). For each ear, pure tone thresholds were tested in the following order: 1000, 2000, 4000, 8000, 6000, 3000, 1500, 750, 500, and 250 Hz.

**Questionnaires.** A modified form of the HSI (Coren and Hakstian, 1992) was administered to each hearing-impaired participant. One item in the original HSI was excluded from this questionnaire because a similar item was included in the LOCATE questionnaire. Participants were instructed to respond to the items as if they were not wearing hearing aids, if they owned any. Significant-other rated indications of SS hearing were obtained using the significant-other version of the modified HSI. Significant others were in-

structed to respond to items as if the hearing-impaired person were not wearing hearing aids, if they owned any. Instructions and questionnaire forms are reproduced in Flamme (2000).

### Understanding in Noise Hearing

**Psychophysical test stimuli.** Target talker and babble stimuli from the CST (Cox *et al.*, 1988) were used. The left babble signal was taken from the compact disk track holding the target sentence. The right babble signal was taken from a separate track of the compact disk. Therefore, the multitalker babble signals were uncorrelated. The talker signal was presented at the reference position at 64 dBA, and each multitalker babble signal was presented at 58 to 59 dBA, with summed babble levels of 62 dBA. These absolute and relative levels were used because they represent appropriate levels for a typical difficult listening situation (Cox *et al.*, 1991; Pearsons *et al.*, 1977). The RMS deviations from a flat 1/3-octave band response were less than 1 dB in the frequency range between 100 and 16,000 Hz.

**Psychophysical test apparatus.** The CST talker and babble signals were played from a CD-ROM and amplified. The CST talker signal was presented from the loudspeaker at 0-degree azimuth and elevation, 1 m from the center of the listener's head. The CST multitalker babble signal was presented from loudspeakers at  $\pm 90$ -degrees azimuths. Note that these loudspeakers also were used in the psychophysical localization test. To control for deviations from a flat frequency response caused by the loudspeakers and room reverberations, CST talker and babble stimuli were filtered using separate channels of Yamaha G-Q1031BII or Q2031A graphic equalizers.

**Psychophysical test procedure.** Testing was preceded by two to three practice passages. During testing, listeners were presented with one set of 12 CST passages, which resulted in 300 scored words. Scores for the CST were calculated, transformed into rau (Studebaker, 1985) and recorded using the CST v.6 software program.

**Questionnaires.** Self-report tests of UN hearing were obtained using the unaided portion of the PHAB, hereafter referred to as the Profile of Hearing Problems (PHP). Instructions asked hearing-impaired listeners to estimate their unaided performance in each situation. Respondents were aware that reversed items were included in the questionnaire.

Significant-other tests of UN hearing were obtained using a modified form of the PHP, hereafter referred to as the PHP-SO. The significant-other form of the PHP was identical to the self report version, except that "I" statements were replaced by "he/she" statements. Instructions stated that participants were to respond to each item as if their significant other were not hearing aids, if they owned any. Significant others were made aware that reversed items were present in the questionnaire.

## Results

This study provides three types of information. First, the data provide estimates of the relationships between the traits of DD hearing, SS hearing, and UN. Second, estimates of the relationships between psychophysical, self-report, and significant-other report measurement methods can be obtained via these data. Finally, one is able to estimate the relative influences of these constructs on a group of nine tests, representing each combination of trait and method. These estimates were made for a group of 49 adult subjects with hearing impairments in unaided conditions.

All tests requiring estimates from significant others were developed for use in this study, and

the LOCATE questionnaire, a self-report test estimating DD hearing, had not previously been administered to unaided listeners with hearing impairments. A complete evaluation of the individual psychometric properties of these four instruments is beyond the scope of this paper, but these data are reported in Flamme (2000).

Analyses of this study's data are reported in three different ways. First, the MTMM correlation matrix is examined. Next, the evidence for convergent and discriminant validity is evaluated in terms of a comparison of the fit of a nested series of confirmatory factor analysis (CFA) models. Finally, the individual coefficients from the MTMM CFA are examined. Results from each analysis approach support similar conclusions.

### Individual Test Results

Descriptive statistics for each of the tests in this study are reported in Table 1. Localization of the click train stimulus is reported as the root mean square (RMS) difference between the actual and judged sound source, in degrees. The tests of DD hearing using self-report and SO-report methods are expressed in the units of the LOCATE and LOCATE-SO response scale, where responses of "almost always," "often," "sometimes," and "almost

**Table 1**  
*Descriptive Statistics*

Test Label	Trait	Method	Mean	SD	Range	K-SZ	Cronbach's alpha
Click localization	DD	Psychophysical	40.7 degrees	28.1	9.5, 108.2	1.69*	Not tested
LOCATE	DD	Self-report	2.6 scale units	0.6	1.3, 4.0	0.57	.93
LOCATE-SO	DD	SO report	2.8 scale units	0.7	1.0, 4.0	0.69	.92
Bilateral 4FAHL	SS	Psychophysical	54.5 dB HL	12.0	27.5, 76.3	0.48	Not tested
HSI	SS	Self-report	3.3 scale units	0.5	2.5, 4.6	0.69	.71
HSI-SO	SS	SO report	3.5 scale units	0.6	2.4, 5.0	0.82	.77
CST	UN	Psychophysical	33.0 rau	32.6	-20.1, 91.6	0.53	Not tested
PHP BN	UN	Self-report	62.2 percent problems	17.2	12.3, 90.8	0.73	.91
PHP-SO BN and RV	UN	SO report	63.7 percent problems	18.0	35.1, 93.1	0.73	.95

N = 49; \*p < .05.

never” were assigned values of 1, 2, 3, and 4, respectively. Thus, mean responses between the values of 2 and 3 indicate that across the 14 items, subjects tended to report correct localization “often” or “sometimes,” and high numbers represent greater amounts of limitation of localization ability.

The test of SS hearing using the psychophysical method was the pure tone threshold, in dB HL, averaged across the frequencies of 500, 1000, 2000, and 4000 Hz, and also averaged across ears. Self- and SO-report tests were expressed in the units of the HSI and HSI-SO response scale. Unlike the LOCATE and LOCATE-SO, the HSI and HSI-SO questionnaire does not use the same response set for all items. For this reason, it is difficult to interpret mean responses in terms of the response categories, however, higher HSI and HSI-SO scores represent greater amounts of soft sound hearing activity limitations.

The test of UN hearing using the psychophysical method was the connected speech test score, expressed in rationalized arcsine units (rau). For this test, higher scores represent lower levels of performance. The self- and SO-report tests of UN hearing provided scores expressed in percentages of hearing problems experienced in daily life, thus higher scores represent greater amounts of problems.

The normality of the distributions of test scores was evaluated using the Kolmogorov-Smirnov (K-S) Z statistic. Only the click localization task was identified as having a non-normal distribution. Examination of the distribution revealed that the mean of the RMS error scores was slightly skewed in the direction of poorer scores.

The internal consistency reliability of test scores was evaluated using Cronbach’s alpha statistic. By subtracting this statistic from 1.0, one obtains an estimate of the amount of random error in the test scores. Thus, high alpha values are desirable, with values exceeding 0.90 being desirable for making clinical decisions (Nunnally and Bernstein, 1994). Note that internal consistency reliability estimates were not obtained for the psychophysical variables because the necessary information was not recorded during data collection.

#### Multitrait-Multimethod Results

The preceding results show that, in general, each of the nine tests in this study returned scores that

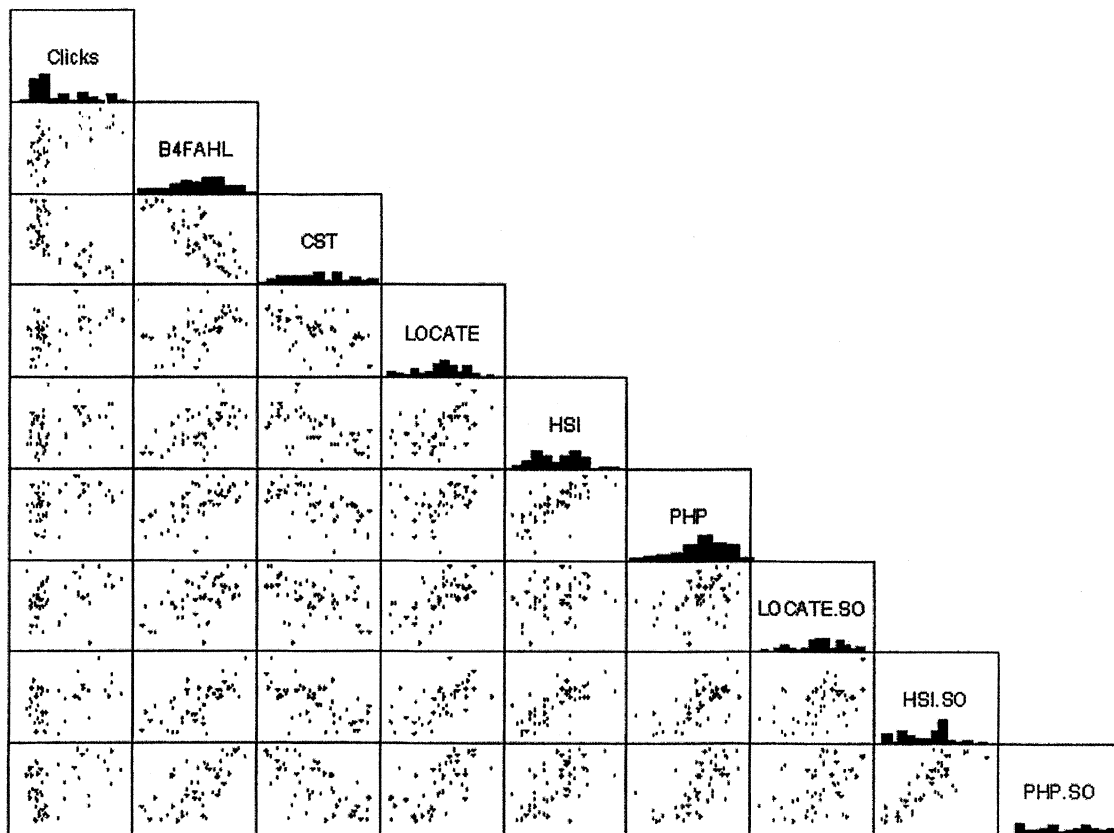
were normally distributed, except in the case of the psychophysical test estimating DD hearing, and all of the measures without relevant prior psychometric data have shown acceptable internal consistency reliability.

Before conducting the Multitrait-Multimethod (MTMM) analyses, a scatterplot matrix of the data was examined for evidence of nonlinear relationships and outliers (Figure 2). Note that the variables in Figure 2 are arranged identically to the MTMM correlation matrix in Table 2, described below. Distributions of each variable are represented along the principal diagonal, in histogram form. In the scatterplot panels of Figure 2, all relationships between variables were observed in the expected directions, with negative relationships between all variables and the CST test, which is scored so that lower values represent poorer performance. Though each of the scatterplots on the off-diagonal of Figure 2 show considerable spread of scores, a clear trend can be seen in each subplot. With the exception of the scatterplots involving the click localization test, the trends are linear. Scatterplots that included the click localization test appeared to have a slight curvilinear trend. In these scatterplots, the addition of a quadratic term to a linear regression model failed to significantly improve the fit of the model; hence all relationships were regarded as linear in the subsequent analyses. It is possible that the apparent curvilinearity is the result of the non-normal distribution of click localization scores.

An objective evaluation of outlying data points implemented in EQS v5.5a did not indicate any data points that had an extreme influence on the results of this study. An overall test of multivariate kurtosis indicated that the data did not significantly deviate from multivariate normality (normalized Mardia’s coefficient =  $-1.502$ ;  $p = 0.129$ ), which means that the data were not shown to violate the assumptions of the inferential statistics used in the confirmatory factor analysis, described below.

#### Multitrait-Multimethod Correlation Matrix

Using SPSS v.8.0, Pearson correlation coefficients were calculated among the nine tests included in this study. Table 2 represents the lower triangle of this multitrait-multimethod correlation matrix. This matrix is organized primarily by measurement method (i.e., PP, SR, SOR), with the traits nested within each level of the method factor.



**Figure 2.** Scatterplot matrix of raw data used in outlier identification. Clicks = click localization test; B4FAHL = Bilateral 4-frequency average hearing level; CST = CST score (rau); LOCATE = LOCATE scale score; HSI = HSI scale score; PHP = unaided PHAB BN scale scores; LOCATE-SO = LOCATE-significant-other scale score; HSI-SO = hearing screening inventory significant-other scale score; PHP-SO = PHAB significant-other scale score.

Therefore, the first row and column represents all correlations associated with the test that combined the psychophysical method and the DD hearing trait, the second row and column represents all correlations associated with the test that combined the psychophysical method and the SS hearing trait, etc. For simplicity, correlations in Table 2 are expressed in absolute values.

The multitrait-multimethod correlation matrix was evaluated using four criteria. First, the validity entries (in boldface) should be significantly different from zero. This criterion provides evidence of convergent validity. All correlations are significantly different from zero; thus, the convergent validity criterion was met, which means that each measure of each trait was shown to be significantly correlated with other measures of the same trait.

The remaining three criteria provide evidence of discriminant validity, which represents the extent to which estimates of different traits provide unique trait-based information. Following the MTMM correlation matrix analysis guidelines of Bagozzi (1993), the discriminant validity criteria are not met. First, within each boxed group of coefficients, the validity entries should be higher than other values lying in the same column and row. Nine of the 36 correlations (25%) fail to meet this criterion. Second, the correlation between two measures of the same trait should be larger than the correlation between two measures of different traits that use the same method. In Table 2, this criterion requires that each of the bolded values should have a greater absolute magnitude than any underlined values. Approximately 64% (25/36) of the correlations

**Table 2**  
*MTMM Correlation Matrix*

Methods		Psychophysical (PP)			Self-Report (SR)			Significant-Other Report (SOR)		
		DD	SS	UN	DD	SS	UN	DD	SS	UN
PP	DD	—								
	SS	.664 <sup>†</sup>	—							
	UN	.712 <sup>†</sup>	.799 <sup>†</sup>	—						
SR	DD	.441 <sup>*</sup>	.582 <sup>‡</sup>	.523 <sup>‡</sup>	—					
	SS	.336 <sup>‡</sup>	.586 <sup>*</sup>	.581 <sup>‡</sup>	.487 <sup>†</sup>	—				
	UN	.430 <sup>‡</sup>	.614 <sup>‡</sup>	.585 <sup>*</sup>	.453 <sup>†</sup>	.701 <sup>†</sup>	—			
SOR	DD	.356 <sup>*</sup>	.464 <sup>‡</sup>	.385 <sup>‡</sup>	.618 <sup>*</sup>	.329 <sup>‡</sup>	.381 <sup>‡</sup>	—		
	SS	.465 <sup>‡</sup>	.680 <sup>*</sup>	.603 <sup>‡</sup>	.617 <sup>‡</sup>	.669 <sup>*</sup>	.648 <sup>‡</sup>	.445 <sup>†</sup>	—	
	UN	.466 <sup>‡</sup>	.677 <sup>‡</sup>	.664 <sup>*</sup>	.610 <sup>‡</sup>	.615 <sup>‡</sup>	.670 <sup>*</sup>	.418 <sup>†</sup>	.805 <sup>†</sup>	—

DD = directional and distance hearing; SS = soft sounds hearing;  
UN = speech understanding in noise.

\*Trait coefficients.

†Method coefficients.

‡Correlations with neither trait nor method in common.

All values are significantly different from zero ( $p < .05$ ).

in Table 2 violate the second criterion. Third, the same pattern of trait interrelationship should be shown in all contiguous groups of small typeface coefficients. These coefficients represent the relationships between measures that share neither trait nor method, so they represent the summed effects of trait covariance and method covariance. If the measures are not greatly influenced by measurement method, these coefficients should have the same rank-order throughout Table 2. For example, in the box in the lower left corner of the table, the lower triangle of three coefficients show the strongest relationship between the estimates of SS and UN hearing, with the next strongest relationship between DD and UN hearing, and the smallest relationship between DD and SS hearing. In the lower triangle of the box immediately to the right, a different rank-order was observed. In this group, the strongest relationship was between DD and SS hearing, the next strongest relationship was between SS and UN hearing, and the weakest relationship was observed between DD and UN hearing. Three rank-order relation-

ships were observed in the triangles representing relationships sharing neither common traits nor common methods. In 50% (3/6) of these triangles, the strongest relationship was noted between SS hearing and UN hearing, followed by the relationship between DD hearing and UN hearing, and finally the weakest relationship was noted between DD hearing and SS hearing. In 33% (2/6) of these triangles, the strongest relationship was noted between DD hearing and SS hearing, followed by the relationship between SS hearing and UN hearing, and finally the weakest relationship was noted between DD hearing and UN hearing. In one of these triangles (17%), the strongest relationship was noted between SS hearing and UN hearing, followed by the relationship between DD hearing and SS hearing, and finally the weakest relationship was noted between DD hearing and UN hearing.

The MTMM correlation matrix (Table 2) gives little evidence of discriminant validity. The discriminant validity criteria were violated in 25%, 64%, and 50% of the cases, although they would

be expected to be violated in no more than 5% of the cases (Bagozzi, 1993). In summary, analysis of the MTMM correlation matrix suggests good evidence of convergent validity, but little evidence of discriminant validity. The finding of good convergent validity means that there is evidence supporting the existence of the traits included in this study. The finding of poor discriminant validity means that the traits included in this study are strongly related to one another, and that although the traits were shown to exist, they are not independent factors and might be completely redundant with one another. So although DD, SS, and UN hearing are conceptually different, the quantitative differences between the traits might be small enough to preclude their estimation in an efficient test protocol. Finally, the MTMM correlation matrix does not offer a mechanism through which individual tests can be evaluated. This type of information is obtained through the CFA that follows.

#### *Confirmatory Factor Analysis*

MTMM CFA provides a more formal evaluation of construct validity than the preceding correlation matrix evaluation. Like exploratory factor analysis (e.g., principal components analysis), CFA fits a structural model to the observed data. However, in the case of a CFA, the investigator is allowed control over the model details (e.g. number of factors, the number of causal influences on a

given test, etc.). This provides a mechanism by which rival hypothetical structures can be empirically compared.

#### MODEL COMPARISONS

The confirmatory factor analysis was performed using the guidelines of Byrne (1994), using EQS v.5.5a. Four nested models were tested. Model 1 is designed such that the covariances between traits and between methods are freely estimated. Model 2 hypothesizes that the observed data was the sole result of method variance, i.e., no traits, but freely correlated methods. Model 3 hypothesizes perfectly correlated traits and freely correlated methods. Method 4 hypothesizes freely correlated traits and perfectly correlated methods.

The overall goodness of fit of each model was evaluated using two statistics, reported in Table 3. The chi-square statistic evaluates the probability of the observed deviations from the specified model's covariance matrix, assuming the specified model is correct. Larger chi-square values indicate a poor fit of the model to the observed data. Thus, a significant chi-square indicates that the model is unlikely to produce the observed data. The comparative fit index (CFI) statistics also were computed. The CFI values range between 0 and 1.0, with a value of 1.0 representing a perfect fit to the data. This index is reported because it provides a relatively accurate estimate of fit in

**Table 3**  
*Summary of Goodness of Fit Indices for MTMM CFA Nested Models*

Model		$\chi^2$	df	Comparative Fit Index
1	Freely correlated traits; Freely correlated methods	5.669	14	1.000
2	No traits; Freely correlated methods	57.486*	26	.884
3	Perfectly correlated traits; Freely correlated methods	10.497	17	1.000
4	Freely correlated traits; Perfectly correlated methods	15.210	17	1.000

\*p < .001.

small samples and because it can be interpreted like the coefficient of determination in regression analyses (Bentler, 1995). The statistics in Table 3 indicate that, although each of the models had high CFI values, Model 1 (i.e., the model with trait correlations that were freely estimated and method correlations that were also freely estimated) provides the best overall fit to the data.

Differences in the goodness of fit indices in Table 3 are used to evaluate the convergent and discriminant validity of the traits in the model. As previously noted, Model 1 provided the best overall fit to the data, however, the fit of Model 1 may not be significantly better than the other models, and these differences are relevant to the validity of the trait and method constructs. Table 4 represents the differences in the goodness of fit indices. Smaller differences in goodness of fit chi-square values indicate that there is little difference in the compared models' abilities to fit the data. The significantly better fit of Model 1 over Model 2 ( $\Delta\chi^2_{(12)} = 51.817$ ;  $p < .001$ ) is evidence of convergent validity for the MTMM CFA model. This is because the model that hypothesized the presence of no traits provides a significantly poorer fit to the observed data than the model that hypothesized that traits were a partial cause of the data.

Discriminant validity, at the model level, was evaluated by the difference tests in the bottom section of Table 4. The test of the discriminant validity of the traits was nonsignificant ( $\Delta\chi^2_{(3)} = 4.828$ ;  $p = .185$ ). This result indicated that a model hypothesizing perfectly correlated traits (i.e., indiscriminable traits) and freely correlated

methods resulted in a practically equal fit to a model with freely correlated, but discriminable, traits and methods. The difference in CFI ( $\Delta\text{CFI} = .000$ ) was consistent with this outcome.

The test of the discriminant validity of the methods was significant ( $\Delta\chi^2_{(3)} = 9.541$ ;  $p = .023$ ). This result indicated that a model hypothesizing perfectly correlated methods (i.e., indiscriminable methods) and freely correlated traits resulted in a significantly poorer fit to the observed data as a model with freely correlated, but discriminable, methods and traits.

#### PARAMETER COMPARISONS

The lack of evidence for discriminant validity of the traits in the goodness-of-fit results suggests that the empirical traits of DD hearing, SS hearing, and UN hearing are very strongly related to one another. To further examine the issues of convergent and discriminant validity, parameter estimates for the best fitting MTMM CFA model (i.e. Model 1) are reported. This model was chosen over the model hypothesizing perfectly correlated traits (i.e. Model 3) because the following data show slight evidence that the correlations between DD and other hearing traits are significantly lower than 1.0, hence the restrictive assumption of perfectly correlated traits seems unjustified. The evaluation of these parameter estimates also provides a way to estimate the relative size of trait and method influences on the tests included in this study.

**Table 4**  
*Differences in Goodness of Fit Indices for MTMM CFA Nested Models*

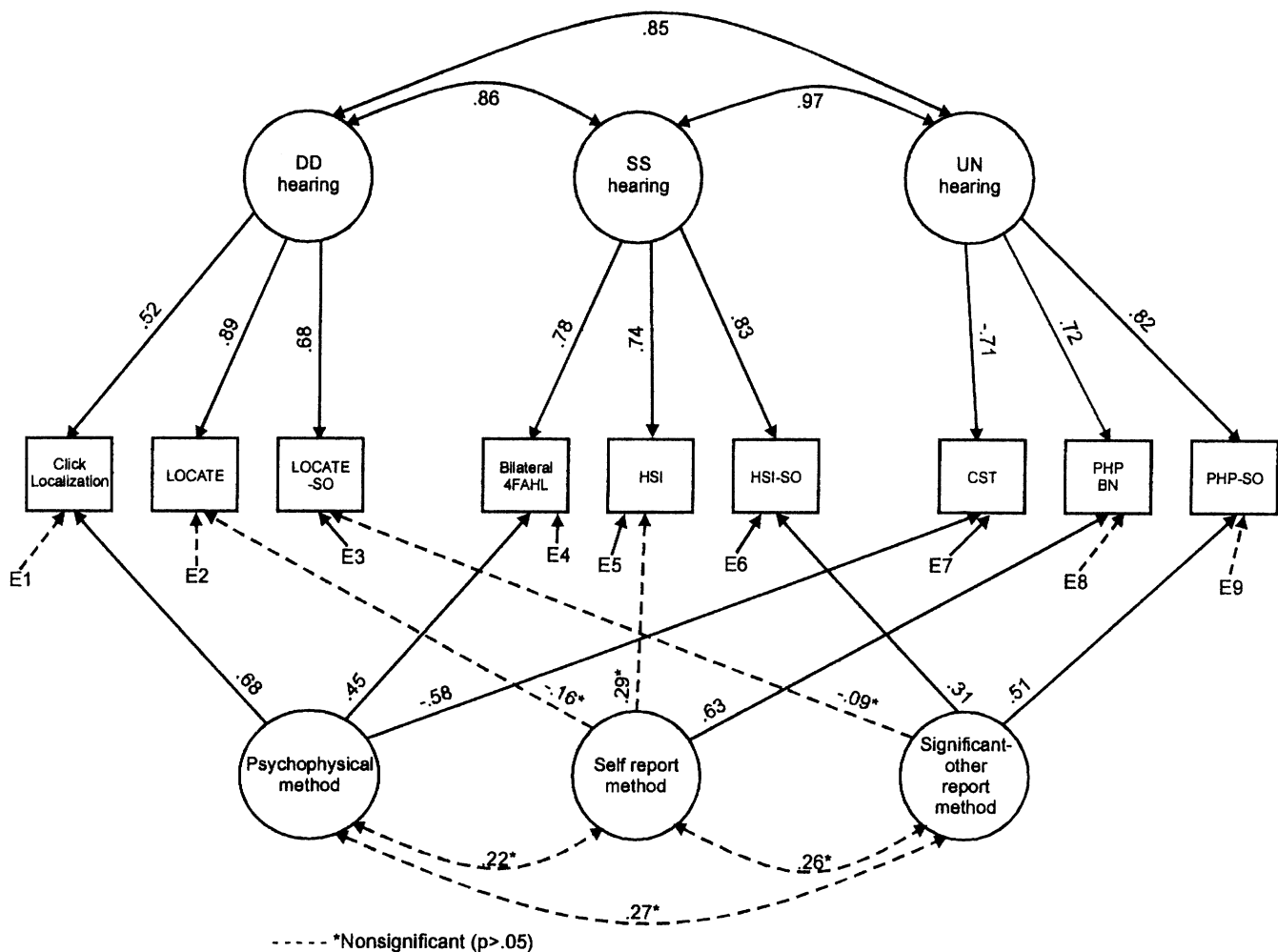
Model Comparisons	Difference ( $\Delta$ ) in		
	$\chi^2$	df	Comparative Fit Index
Test of Convergent Validity			
Model 1 vs Model 2	51.817*	12	.066
Test of Discriminant Validity			
Model 1 vs Model 3 (traits)	4.828	3	.000
Model 1 vs Model 4 (methods)	9.541*	3	.000

\* $p < .05$ ; \* $p < .001$ .

The best-fitting parameter estimates for Model 1 are presented in Figure 3. Scores from the individual measurement operations (i.e., tests) are represented by squares. Circles represent the constructs hypothesized to cause some of the systematic variance in these scores. The constructs at the top of the figure are trait factors; those at the bottom are method factors. Unidirectional arrows represent supposed causal relationships; bidirectional arrows represent associations without a hypothesized causal direction.

Examination of each test (i.e., each square) will show that each test is hypothesized to be

caused by three factors. One of these factors is the substantive trait presumed to be estimated by the test, the second is the measurement method, and the third is a unique/error factor representing all variance that cannot be attributed to either of the other sources. One can interpret each test in the model as a dependent variable in a multiple regression equation, where the coefficients leading to each test represent standardized regression coefficients predicting the outcome of that test. For example, for the click localization test, the form of the regression equation would be:



**Figure 3.** Structure for MTMM CFA Model 1 and best-fitting parameter estimates. Squares represent tests, circles represent trait and method constructs. All solid arrows represent significant relationships. Broken arrows represent nonsignificant relationships. High correlations between traits indicate that the trait components of the tests are strongly related. High values between the constructs and tests indicate that the test is highly loaded with the construct. See text for further details.

$$Y = .52 \times DD + .68 \times PM + E_1$$

where  $Y$  represents the outcome of the click localization test,  $DD$  represents the subject's performance on the direction and distance hearing trait factor,  $PM$  represents the subject's reaction to psychophysical methods, and  $E$  represents an error term that includes random error variance and all other variance in the test that cannot be attributed to trait nor method. Note that the coefficients in this regression equation, when squared, will be equal to 1.0; that is, the model structure will partition the variance of the test into trait, method, and unique/error components. In factor analysis, these coefficients are called factor loadings. In cases where the causal coefficient was not significant at the  $p < .05$  level, the arrows are represented as broken lines. Therefore, all continuous arrows represent significant associations.

Table 5 represents the factor loadings and communalities (calculated as the sum of squared factor loadings for a given test) from Model 1. Significant factor loadings for traits are evidence of convergent validity for that test. Table 5 shows that trait-based factor loadings ranged between .894 and .523, and all were significant at the  $p <$

.001 level. Significant method factor loadings indicate that method has a significant impact on the outcome of a test. The significance of each factor loading is represented in Table 5.

The significant trait-based loadings indicate that each test provided evidence of convergent validity. Thus, each test score provided a valid estimate of its respective trait (i.e., all of the supposed tests of DD hearing are significantly related to the DD hearing empirical trait, etc.). With the exception of the LOCATE, LOCATE-SO, and the HSI, all tests are significantly influenced by method variance, though the degree of saturation varied across tests. This finding suggests that the majority of the tests in this study were significantly impacted by intrasubject response biases associated with the measurement method.

Note that the communality values from Model 1 were quite high (mean communality = .770). These suggest that the parameter values emerging from the confirmatory factor analysis can be expected to show a high concordance with population values (MacCallum *et al.*, 1999), despite the relatively small sample size in this study.

Discriminant validity is estimated via the correlations between trait factors. Large between-

**Table 5**  
*MTMM CFA Factor Loadings and Communalities*

Variables	Traits			Methods			Communalities
	DD	SS	UN	Psychophysical Method	Self-Report Method	SO Report Method	
PP LOC	.523*			.680*			.736
LOCATE	.894*				-.161		.825
LOC-SO	.675*					-.088	.463
4FAHL		.776*		.452†			.807
HSI		.737*			.286		.625
HSI-SO		.828*				.305†	.779
CST			-.709*	-.576*			.835
PHP			.722*		.635‡		.925
PHP-SO			.817*			.514†	.932

PP LOC = psychophysical localization task; LOC-SO = LOCATE-SO; DD = direction and distance hearing; SS = soft sounds hearing; UN = understanding in noise hearing.

\* $p < .001$ ; † $p < .01$ ; ‡ $p < .05$ .

factor correlations indicate that small amounts of unique information can be obtained by estimating a person's performance on one trait, given an accurate estimate of the listener's performance on the other trait. Hence, the factors are not easily discriminable, and clinicians or clinical researchers wishing to obtain an estimate of performance on one of these traits that is not redundant with the other traits should plan to use a test with an abnormally high proportion of trait variance. Table 6 represents the lower triangle of the between-factor correlation matrix. All correlations among traits were strong and significantly different from zero below the  $p = .00001$  level.

Although the significance of a between-trait correlation provides a rough estimate of discriminant validity, the confidence interval surrounding the observed correlation is of greater interest. Two traits can have a nonzero relationship, yet still be sufficiently different to justify estimation in a test protocol.

These confidence intervals were estimated through a small simulation study. In this approach, 100 data sets of the same size were synthesized, combining systematic variance and random error using the approach of MacCallum and associates (1999) using the population correlation matrix estimated by Model 1. Each of the data sets was analyzed using EQS 5.5a, the be-

tween-factor correlations were recorded, and the 2.5% and 97.5% points on the distribution of each correlation were identified across these 100 estimates. Using the simulation study data, the 95% confidence interval surrounding the correlation between DD hearing and SS hearing was 0.72 to 0.91. The 95% confidence interval surrounding the correlation between DD hearing and UN hearing was 0.53 to 0.92. The 95% confidence interval surrounding the correlation between UN hearing and SS hearing ranged between 0.91 and 1.00.

The confidence intervals estimated via these two methods indicate that a strong relationship exists between the trait factors. In the case of the relationship between SS and UN hearing, the correlation between the traits is not significantly different from a perfect correlation, once the effects of systematic and unsystematic irrelevant variance have been controlled. The simulation study-based confidence intervals indicate that DD hearing may contain a small amount of information that would not be obtained via tests returning perfect estimates of the other traits in this study. Method factor correlations were not significantly different from zero, indicating that psychophysical, self-report, and significant-other report methods impact scores in independent ways.

**Table 6**  
*MTMM Factor Correlation Matrix*

Correlate	DD	SS	UN	Psycho- physical Method (PP)	Self- Report Method (SR)	SO Report Method (SOR)
DD	1.000					
SS	.849*	1.000				
UN	.855*	.967*	1.000			
PP	—	—	—	1.000		
SR	—	—	—	.220†	1.000	
SOR	—	—	—	.273†	.258†	1.000

— = Correlations fixed at zero; DD = direction and distance hearing; SS = soft sounds hearing; UN = understanding in noise hearing.

\* $p < .000001$ ; † $p > .05$ .

Another way to look at the trait and method influences on the tests included in this study is by examining the size of the trait, method, and error variance components for each test. Table 7 reports the factor loadings for each test, squared to represent variance components. Note that for all but psychophysical estimate of DD hearing, the trait accounts for the majority of each test's variance. Note further that a substantial amount of variance in many of the tests is not trait-related. This finding is of considerable concern because test results are typically interpreted as unbiased estimates of a listener's performance on the trait. Particularly for the tests containing large amounts of method variance, these estimates are likely to be biased upward or downward, depending on the listener's reaction to the measurement operation.

#### Exploratory Factor Analysis

The correlations between DD hearing and other traits were substantially higher than those observed in prior studies. This difference could either be due to a difference in the analysis approaches, or fundamental differences in the subject population sampled in this study. To distinguish between these possible causes, the current data were compared with the data obtained by

Kramer and associates (1995) via an exploratory principal components analysis. Three factors were rotated using a direct oblimin criterion ( $\delta = 0$ ). Based on the pattern of factor loadings, these factors appeared to represent DD hearing, a merged SS/UN hearing factor, and a factor representing the psychophysical method. The results showed that the DD trait factor and the merged SS/UN factor had a between-factor correlation of approximately .48, which is quite similar to the correlation observed in Kramer and associates ( $\rho = .43$ ). The difference between the correlations observed via confirmatory and exploratory factor analysis results seem to be a function of the different causal models hypothesized in each approach (Long, 1983). The implications of these differences are discussed below.

#### Discussion

The goals of this study were to examine the relationship among three hearing traits, direction and distance (DD) hearing, soft sounds (SS) hearing, and understanding in noise (UN) hearing, while also estimating the amounts of trait-related, method-related, and other influences on tests de-

**Table 7**  
*Proportions of Trait, Method, and Unique/Error Variance from the MTMM CFA*

Variable	Trait	Method	Unique/Error
Psychophysical DD	.27	.46	.26
LOCATE	.80	.03	.18
LOCATE-SO	.46	.01	.54
4FAHL	.60	.20	.19
HSI	.54	.08	.38
HSI-SO	.69	.09	.22
CST	.50	.33	.17
PHP	.52	.40	.08
PHP-SO	.67	.26	.07

signed to return estimates of these traits. The results of this study suggest that these traits are very strongly related to one another, which means that the trait-based intersubject differences identified by tests of unaided SS hearing are largely the same as those identified by unaided tests of DD hearing and UN hearing. In addition, the results suggest that the method-based variance in psychophysical, self-report, and significant-other report instruments are unrelated.

The strong relationships among traits do not mean that scores from tests intended to estimate those traits are also redundant. It only means that the trait-based components of the test scores are redundant. An examination of the amount of scatter in Figure 3 reveals that the relative scores across tests from a given listener often were quite different (i.e. a listener could have had a relatively high score on one test of a trait, but a relatively low score on another test of the same trait). This difference is of special concern for clinicians. Clinicians evaluate one listener at a time, and must use limited data to decide what is represented by a difference between test scores. The evaluations of discriminant validity in this study suggest that differences between tests estimating one trait and tests estimating another trait are not likely trait-based. For unaided adults with sensorineural hearing impairment, disagreements between scores from these tests are most likely to result from conflicting method biases and/or excessive error variance in at least one test.

The current results address an apparent conflict in prior studies. Noble and associates (1997) found no significant correlation between psychophysical tests of DD hearing and UN hearing after psychophysically measured SS hearing was controlled. Noble and associates (1995) observed a significant partial correlation between self-report tests of DD hearing and UN hearing, after psychophysically measured SS hearing was controlled. The current results suggest that a likely cause for this difference is that the variables used in the Noble and associates (1997) partial correlation shared a common method, while the variables in the Noble and associates (1995) partial correlation did not share a common method. Assuming that the tests in those studies were influenced by method variance, the Noble and associates (1995) significant partial correlation could have represented method covariance, while the corresponding correlation in Noble and associates (1997) partialled out both trait and method covariance.

The current study shows substantially stronger relationships between DD and UN hearing than previously observed (e.g., Kramer *et al.*, 1995; Lutman *et al.*, 1987; Ringdahl *et al.*, 1998). Different analysis methods are the probable cause of this difference. All previous studies in this area used exploratory principal components analysis. The exploratory factor analysis model facilitates weaker relationships between factors because all observed variables are modeled to be jointly caused by all common factors or components (Long, 1983). In the current study, tests were hypothesized to be impacted by only one trait, which is similar to how tests are scored and interpreted. Lutman and co-workers (1987) and Ringdahl and associates (1998) used varimax factor rotation, which imposes a hypothesis of uncorrelated factors on the data. Varimax rotation tends to overcomplicate simple causal structures; it was designed to eliminate general factors (see Nunnally and Bernstein 1994, p. 506). Kramer and associates (1995) used an oblique rotation of the factors, which found only a moderate correlation between self report tests of the DD hearing and UN hearing factors. When the current data were analyzed similarly to methods of Kramer and associates (1995), the correlation between the DD hearing factor and the merged SS/UN hearing factor was quite close to the factor correlations observed in their study ( $\rho = .48$  vs  $\rho = .43$ , respectively).

---

#### Pure Tone Thresholds as an Estimate of Hearing Disability

---

Pure tone thresholds have long been considered insufficient indicators of hearing problems in daily life (Demorest and Walden, 1984; Erdman, 1994; Erdman and Demorest, 1998; Gatehouse, 1994; Hallberg and Carlsson, 1991; High *et al.*, 1964; Karlsson and Rosenhall, 1998; Kramer *et al.*, 1996; Marcus-Bernstein, 1986; Newman *et al.*, 1990; Noble and Atherley, 1970; Schow and Nerbonne, 1980; Speaks *et al.*, 1970; Swan and Gatehouse, 1990). The most common interpretation of this weaker-than-expected relationship is that pure tone threshold procedures sample a different trait than what dominates hearing disability and handicap in daily life. Therefore, SS hearing is expected to be empirically different from UN and other types of hearing. The current results indicate that SS hearing is strongly related

to DD and UN hearing. The principal cause for finding moderate correlations between self-reported and psychophysical measures of hearing appears to be due to the influence of measurement method. The correlations between tests estimating DD, SS, and UN hearing appear to be limited by divergent methods, rather than divergent traits. This means that clinicians observing a discrepancy between scores on auditory tests should not interpret the differences as an indication of inherently different trait abilities. Instead, the clinician should look to the other two components of each test (i.e., method variance and error variance) for the cause of the discrepancy. One could perhaps discriminate between the method and error variance components by determining whether the client also shows unexpectedly good or poor performance on other pairs of measures employing the same method combination. If atypical score combinations are observed with these additional measures, method variance can be considered the most likely cause of this discrepancy. The tester could then evaluate possible reasons why measures employing a certain method might be biased. In the case of psychophysical measures, improper calibration, misunderstood instructions, and high levels of certainty (on the listener's part) required before making a response (i.e., a high response criterion) could be considered potential causes for biased psychophysical test results. In self report measures, a propensity toward minimizing or maximizing problems, the client's activity level and lifestyle, and aberrant response patterns (e.g., avoidance of or attraction to the center of the response scale, etc.) could be considered potential causes for biased results. Some of these potential causes for bias could be important to a client's rehabilitative plan, while others are nuisance factors.

Based on the factor intercorrelations from the MTMM CFA, SS, and UN hearing factors were essentially redundant in this sample. The observed correlation between these factors was 0.97, and the lower bound of the confidence interval surrounding this correlation was approximately 0.90. But it remains unclear whether, at the group level, it is more desirable to obtain the parsimony of assuming that DD, SS, and UN hearing are essentially redundant, or if it would be better to consider DD hearing as a strongly related yet slightly different trait. Future work is needed to replicate these findings and examine the implications of issue.

Although the current data indicate very strong relationships between traits, it is incorrect to infer that the administration of a single test of one of these traits would provide an adequate estimate of a listener's underlying hearing abilities. Each test included in this study was shown to be at least partially flawed because of (a) substantial method variance, (b) substantial error variance, or (c) substantial amounts of both. Table 7 shows that, on average, less than 60% of the variance in the tests was due to the trait, which means that a great deal of the intersubject differences in test scores are unrelated to the trait that the tester was intending to estimate.

#### How Can Method Variance be Interpreted?

Mathematically, method variance is a pattern of intersubject differences that is associated with the measurement method, but not the trait. For example, some participants might expect that giving the same response to all items in a questionnaire could be considered undesirable to the tester, and therefore they might avoid such a response pattern. This would be an example of the good subject effect described by Rosenthal and Rosnow (1991). In addition, participants might respond to the first items of a given type with relative uncertainty or randomness, but then bias later judgments to be consistent with previous items that they consider similar. A similar process can be hypothesized in psychophysical tests, where different listeners may set different criteria for responses (e.g., different loudness levels required to before reporting hearing the tone), and then keep this criterion throughout the test. Also, different listeners may make different assumptions about the test procedure and rely on those assumptions during the test. For example, in the psychophysical estimate of DD hearing, subjects could have formed early assumptions about the order of signal presentation across loudspeakers, in the absence of helpful auditory cues, matched their responses to these assumptions. This would be especially likely with those who found the task very difficult, because the amount of response bias has been shown to increase with task difficulty (Lorenzi *et al.*, 1999a; Lorenzi *et al.*, 1999b).

Self-report method variance components might reflect important personal adjustment and rehabilitative factors, regardless of whether they are related to the listener's performance on the auditory trait. Listeners who are rarely exposed

to difficult listening situations (e.g., those who rarely need to understand in noisy situations) might not report as much difficulty in noisy situations as listeners with the same underlying ability but who are more often exposed to difficult listening situations. Note that this type of method variance previously described could be important to a client's rehabilitative plan. Though it is unrelated to the client's inherent hearing abilities, this type of method variance could provide meaningful information about the listener's function in daily life, in the sense that it relates to the frequency of problems rather than the magnitude of problems in a given listening situation.

Many of the tests in this study were shown to be influenced by measurement method. This suggests that, in the absence of direct evidence to the contrary, auditory test results should be considered likely to contain considerable non-trait variance. Two thirds of the tests in this study had significant method factor loadings; the remaining tests had considerable proportions of unique/error variance. Psychophysical tests were the most saturated with method variance. The smallest psychophysical method loading was observed with 4FAHL, with 20% of the variance in 4FAHL was associated with the measurement method. In the psychophysical localization test, method variance accounted for a greater proportion of variance than trait variance (46% and 27%, respectively). This outcome could have been due to the difficulty of the tasks in the psychophysical test of DD hearing. The stimulus parameters for the psychophysical tests in this study were configured to identify, for the average subject, the boundary between hearing and not hearing. It is possible that this region of performance is especially open to method biases.

The observation of uncorrelated method factors suggests that the biases that a person uses with one method are unrelated to those used with another method. This means that if a person's psychophysical test of SS hearing appears worse because of a biasing strategy, that person is not more likely to also bias self-report tests of SS hearing toward worse scores.

To our knowledge, no prior MTMM studies have been identified in the audiology literature, although method effects have been mentioned in some studies (Demorest and Walden, 1984; Erdman and Demorest, 1998; Walden *et al.*, 1984). The results of the current study indicate that method variance can impact many audiolog-

ic measures. The implication of this finding is that method covariance should be considered a potential cause for any observed covariance between measures using identical methods. For example, a psychophysical estimate of SS hearing might be correlated with a psychophysical estimate of sound annoyance because of a shared method, rather than a relationship between the traits.

A clinician could perhaps discriminate between the method and error variance components by determining whether the client also shows unexpectedly good or poor performance on other pairs of measures using the same combination of methods. For example, if the listener shows unexpectedly high amounts of unaided hearing problems on the background noise subscale of the PHAB, given their average pure tone threshold, a clinician could administer another questionnaire (e.g., the LOCATE or HSI), and another psychophysical test (e.g., the CST) and examine whether the same type and magnitude of difference is observed. If similar atypical score combinations are observed with these additional measures, method variance can be considered the most likely cause of this discrepancy. The clinician could then evaluate possible reasons why measures employing a certain method might be biased. If the atypical response is limited to only one test, the most likely cause would appear to be measurement error. Recent problems with a key situation, misunderstood instructions, improper calibration, and transient attention problems could be considered potential causes for unexpected scores on a single test. Note again that some of these factors could be important to a client's rehabilitative plan, but are not directly related to the client's hearing ability. Finally, it would be desirable to add another observation method (e.g., the reports of significant others) in cases of score discrepancies. The scores for each indicator could be integrated to create a single score.

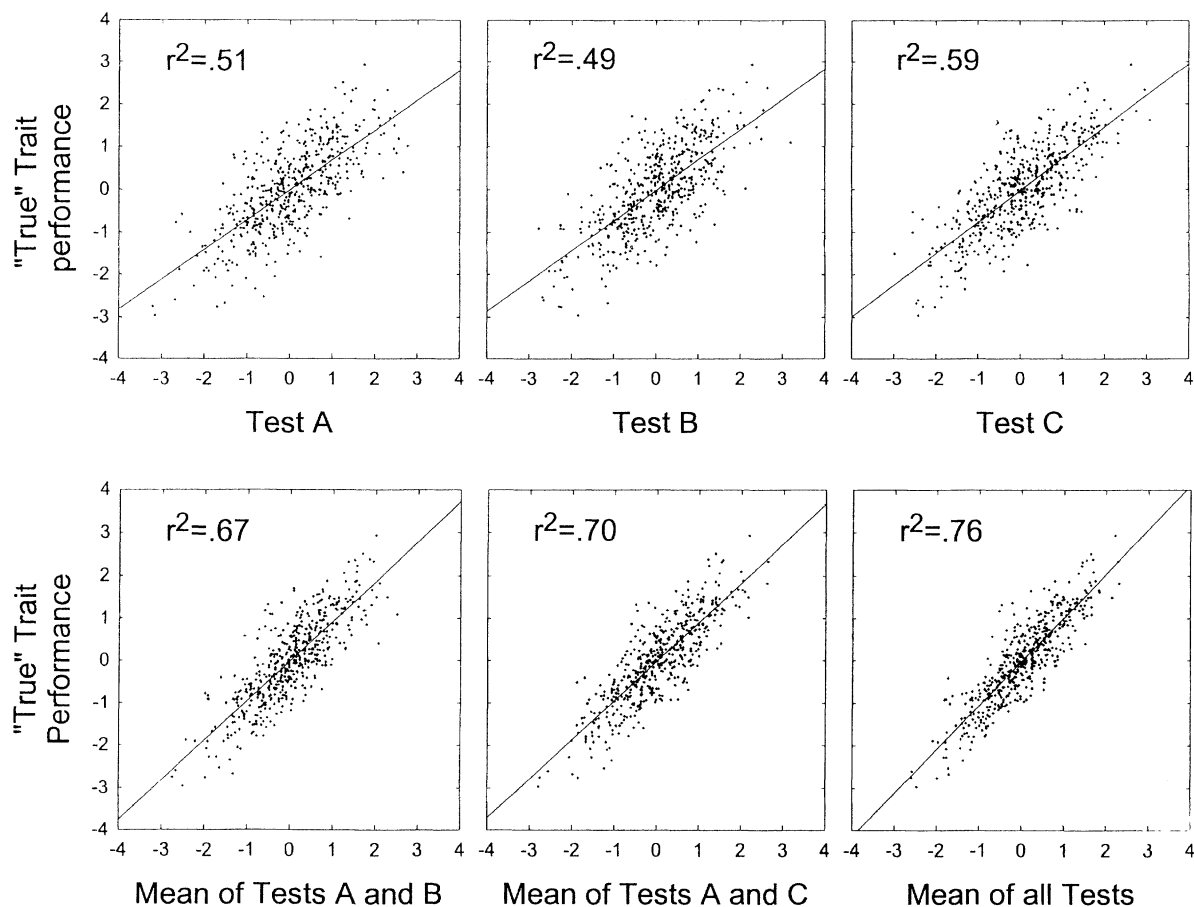
In cases where testers are required to use tests with relatively low amounts of trait variance, the current results suggest that it is reasonable, even desirable, to combine scores obtained from psychophysical, self-report, and significant-other report indicators of unaided DD hearing, SS hearing, and UN hearing. By combining scores from instruments employing different methods, it is possible to reduce the influence of method variance. Such an index of unaided performance

would allow the clinician to obtain a purer estimate of the listener's hearing function.

To illustrate the benefit of combining scores across measurement method, a statistical simulation model was used. The luxury of using a model in this demonstration is that, although it is impossible to know the true performance of a given listener in actual conditions, the results of the simulated tests can be compared against prespecified true trait abilities (i.e., the trait factor). Using trait, method, and error variance proportions similar to those found in the current study, simulated scores were defined for each of three tests. The simulated scores from these tests represent the combined influences of trait, method,

and random error. Across simulated tests, a single trait factor was used to generate simulated scores, but method and error influences were unrelated to each other and to the trait.

In the simulation, better estimates of trait scores were obtained as more tests were combined. These results are represented in Figure 4. The vertical axis in each panel of Figure 4 represents the true performance on the trait, and the horizontal axis represents the simulated test's estimate of true performance. Thus, the spread of data points is an indication of the inaccuracy of the test as an estimator of the trait. To facilitate the combination of scores across tests, all scores are reported in standard (z) score units.



**Figure 4.** Effect of combined scores on the accuracy of trait performance estimates. In cases where all tests have substantial method variance, the combination of scores across tests with unrelated method variance can result in better estimates of listener performance on the trait.

A considerable spread of scores around the true trait value is found in all panels of Figure 4. The poorest estimates of trait performance were observed for single tests, with decreasing errors of estimation as tests were combined. For test A, the proportion of between-score differences that are due to differences in the trait is 0.51. Thus, slightly more than one-half of the variance in test A relates to the quantity that it was intended to measure. The remaining intersubject differences in test A were due to method variance and error variance. Similar results were observed for test B ( $r^2 = .49$ ), and slightly improved results were observed for test C ( $r^2 = .59$ ). The bottom panels of Figure 4 show that the simple combination of test scores across unrelated measurement methods (e.g., psychophysical, self-report, and SO report) results in improved estimates of trait performance. Clinicians will note that combining results across tests is intuitively appealing and is informally applied in clinical settings. However, the improved accuracy of combined estimates relies on the combination of tests having unrelated method influences. The combination of separate scores on parallel forms of the same test would only reduce the influence of the error component of the test score, and would reinforce the influence of the method component.

It is desirable to separate the impacts of trait and method factors because it gives the tester a better characterization of a listener's hearing function. However, the desirability of an unbiased estimate of performance on a trait does not imply that method variance might not also contain important information. Method variance might be related to non-auditory traits that are powerful determinants of a person's participation restrictions. Personality, adjustment, coping, and other important psychosocial factors might be represented in method variance. As noted below, further work in this area is needed.

---

#### Method Variance and Hearing Aid Comparisons

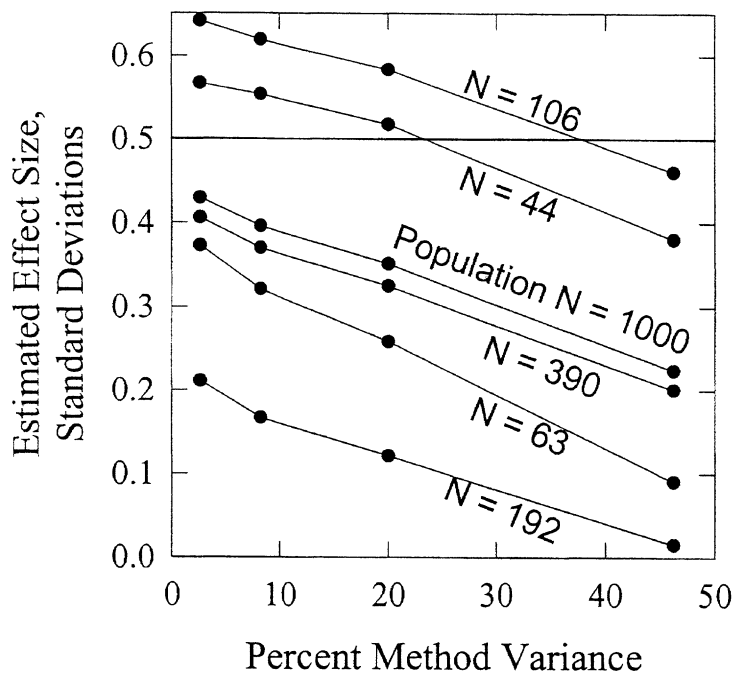
---

The data from the current study have shown that the results of many auditory tests are heavily influenced by non-auditory factors. These non-auditory influences do more than complicate the assessment of a person's hearing performance. They also reduce the likelihood that true differences

between hearing aids (or other treatments) will be detected in clinical trials and other treatment comparison studies.

A study's ability to detect true underlying differences between treatments (e.g. hearing aid X vs hearing aid Y) is the domain of statistical power analysis. The equations of statistical power analysis typically do not account for differences in the proportions of irrelevant variance in the outcome variable (i.e., the test chosen to indicate whether one treatment option is superior to another). This is undesirable because the results of the current study indicate that a relatively large proportion of between-subject differences in auditory tests are not likely to be changed by a treatment, because treatments typically are not designed to alter a listener's response to the measurement situation. In addition, between-subject differences related to measurement method are systematic and not likely to average out in the manner presumed by typical power equations. This means that substantial treatment differences are likely to be missed if the outcome variable is heavily loaded with irrelevant variance. Figure 5 illustrates this effect. The data represented in Figure 5 are the result of a (second) simulation study, where a large treatment difference (i.e., effect size) of 0.5 standard deviation units was applied to the trait scores of 1000 simulated observations. Simulated test scores were generated across varying levels of method variance and sample sizes.

The most obvious feature of the data in Figure 5 is the decrease in the observed effect size as the test scores became more heavily loaded with method variance. This means that observed differences between treatments are biased toward smaller values in tests containing larger amounts of method variance, which implies that even when study sample sizes are determined using conventional equations, the ability of a study to detect true differences between treatments is reduced when tests contain large amounts of method variance. Note that the same trend was observed across all samples of the simulated population, which suggests that the trend is a general impact of method variance on treatment comparisons. The results of this simulation study suggest that in clinical trials and other studies comparing the impacts of treatment options, study designers should attempt to use test or other outcome variables known to have small amounts of method variance.



**Figure 5.** Impact of method variance on estimates of treatment effects. True effect size = 0.5 standard deviations. Increased amounts of method variance in test scores reduces the observed size of effects, reducing the likelihood of detecting better treatments in clinical trials.

#### Future Research Needs

The greatest future research need pertaining to the current study is the need for cross validation with a larger sample. The results of the simulation study suggest that the overall conclusions of this study would not change if the study were replicated. However, replication with a larger sample will likely provide a narrower confidence interval surrounding parameter estimates (e.g., factor correlation coefficients) and also allow the exploration of post hoc models. For example, it would have been desirable to test a model where each of the three tests of DD hearing was jointly determined by DD hearing and SS hearing. However, this model could not have a significantly better fit to the sample data than Model 1 (freely correlated traits and methods), because the maximum chi-square difference would have been less than 6, and the  $p < 0.05$  criterion chi-square with 3 df is approximately 8. A larger overall chi-square statistic would be expected with a larger sample, which would allow the examination of rival models.

Further work is needed to determine whether similar trait relationships are noted under aided conditions. The current study suggests that, on average, the conditions conveying SS hearing also convey UN hearing and, to a slightly lesser extent, DD hearing. However, the relationships might be different under aided conditions. DD hearing, SS hearing, and UN hearing might not be equally impacted by hearing aids. Hearing aids improve SS hearing by amplifying environmental sounds above the listener's unaided ability. DD hearing is partially based on binaural cues, which appear to be degraded by hearing impairment (Lorenzi *et al.*, 1999b). UN hearing might consist of two factors, attenuation and distortion (Plomp, 1978). In an unaided context, these factors are confounded. Hearing aids might improve SS hearing, but leave DD and UN hearing relatively unaffected.

It would be helpful to know more about the amounts of method effects in other tests and/or acoustical conditions. In addition, the amounts of method variance present in some of the current

study's tests warrants further work on the nature, measurement, correlates, and control of method variance. Acquiescence, social desirability distortions, demand characteristics (Rosenthal and Rosnow, 1991), and personality factors might be represented in method variance components. There could be useful information in an understanding of a client's method-based biases. Method variance might be associated with other factors, information about which could impact a client's rehabilitative plan.

### Summary

- In an unaided context, and in spite of obvious conceptual differences, direction and distance hearing, soft sounds hearing, and speech understanding in noise hearing are so strongly interrelated that accurate measures of a listener's particular abilities on one of these traits would be difficult to obtain, given prior knowledge of performance on at least one of the others.
- The tests that are commonly used to estimate a listener's hearing performance do not provide sufficiently accurate and unbiased information. Most auditory tests are heavily influenced by method-related factors that are not impacted by the trait expected to be measured by the test.
- Method-related variance in test scores is not related to the trait of interest, but might still provide useful information about the person being tested.
- If a tester must use a set of tests with small amounts of trait variance, he/she can obtain better estimates of a person's performance on the trait by:
  - including tests that use very different measurement methods (e.g., psychophysical, self-report, and significant-other report)
  - converting the scores from these tests into the same metric (e.g., standardized scores) and
  - combining the scores into an overall score.

Although this is informally done in most clinical situations, formal procedures and norms should be developed.

- In clinical trials, outcome measures with large amounts of method variance are less likely to identify treatment options that are indeed beneficial. Method variance is systematic (i.e., an individual's bias associated with a measurement method can be expected to be stable over time and consistently different from people with other method biases. Because of this, the method variance component of a person's test score cannot be expected to average out over repeated samples, nor can it be expected to be changed by a treatment designed to change performance on a trait. Observable differences between treatments are reduced by method variance in outcome measures.

### Acknowledgments

The author thanks Robyn Cox and Will Shadish for their consultation and assistance with this project, and Ruth Bentler, Martyn Hyde, and an anonymous reviewer for their helpful comments on an earlier version of this manuscript. Chris Bentler and Jason Ruyle also deserve recognition for their help entering the simulation study data.

### References

- American National Standards Institute (1998, August). American National Standard Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms, S3.1-199x (draft). New York: American Institute of Physics.
- American National Standards Institute (1986). Methods for Manual Pure-Tone Threshold Audiometry, S3.21-1986. New York: American Institute of Physics.
- Arrindell WA, van der Ende J. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *App Psychol Meas* 9:165-178, 1985.
- Bagozzi RP. Assessing construct validity in personality research: Applications to measures of self-esteem. *J Res Pers* 27:49-87, 1993.
- Barrett PT, Kline P. The observation to variable ratio in factor analysis. *Pers Study Group Behav* 1:23-33, 1981.
- Bentler PM. EQS Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc., 1995.
- Bentler RA, Kramer SE. Guidelines for choosing a self-report outcome measure. *Ear Hear* 21:37S-49S, 2000.

- Blauert J. Spatial hearing: The psychophysics of human sound localization (3rd ed.). Cambridge, MA: MIT Press, 1997.
- Byrne BM. Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications and programming. Thousand Oaks, CA: Sage Publications, Inc., 1994.
- Cattell RB. The Scientific Use of Factor Analysis. New York, NY: Plenum, 1978.
- Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56:81-105, 1959.
- Chmiel R, Jerger J. Some factors affecting assessment of hearing handicap in the elderly. J Am Acad Audiol 4:249-257, 1993.
- Coren S, Hakstian AR. The development and cross-validation of a self-report inventory to assess pure-tone threshold hearing sensitivity. J Speech Hear Res 35:921-928, 1992.
- Cox RM, Alexander GC, Gilmore C. Intelligibility of average talkers in typical listening environments. J Acoust Soc Am 81:1598-1608, 1987a.
- Cox RM, Alexander GC, Gilmore C. Development of the Connected Speech Test (CST). Ear Hear 8 (suppl): 119S-126S, 1987b.
- Cox RM, Alexander GC, Rivera I. Accuracy of audiometric test room simulations of three real-world listening environments. J Acoust Soc Am 90:764-772, 1991.
- Cox RM, Alexander GC, Gilmore C, Pusakulich KM. Use of the Connected Speech Test (CST) with hearing-impaired listeners. Ear Hear 9:198-207, 1988.
- Cox RM, Alexander GC, Gilmore C, Pusakulich KM. The Connected Speech Test version 3: Audiovisual administration. Ear Hear 10:29-32, 1989.
- Cox RM, Gilmore C. Development of the Profile of Hearing Aid Performance (PHAP). J Speech Hear Res 33:343-357, 1990.
- Cox RM, Gilmore CG, Alexander GC. Comparison of two questionnaires for patient-assessed hearing aid benefit. J Am Acad Audiol 2:134-145, 1991.
- Cronbach LJ. Response sets and test validity. Educ Psychol Meas 6:475-494, 1946.
- Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 52:281-302, 1955.
- Demorest ME, Walden BE. Psychometric principles in the selection, interpretation, and evaluation of communication self assessment inventories. J Speech Hear Disord 49:226-240, 1984.
- Erdman SA. Self-assessment: From research focus to research tool. J Acad Rehabil Audiol 27:67-90, 1994.
- Erdman SA, Demorest ME. Adjustment to hearing impairment II: Audiological and demographic correlates. J Speech Lang Hear Res 41:123-136, 1998.
- Flamme GA, Cox RM, Alexander GC, Gray G. Localization disabilities in real-world situations. Poster presented at the 1999 American Academy of Audiology Annual Convention, Miami FL, 1999.
- Flamme GA. The relationships between direction and distance hearing and other auditory traits: A multitrait-multimethod evaluation. University Microfilms International. Publication number AAT 9967035, 2000.
- Gatehouse S. Components and determinants of hearing aid benefit. Ear Hear 15:30-49, 1994.
- Gatehouse S. Self-report outcome measures for adult hearing aid services, some uses, users, and options. Trends Amplification 5:00-00, 2001.
- Good MD, Gilkey RH. Sound localization in noise: The effect of signal-to-noise ratio. J Acoust Soc Am 99:1108-1117, 1996.
- Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. Psychol Bull 103:265-275, 1988.
- Hallberg LRM, Carlsson SG. Hearing impairment, coping and perceived hearing handicap in middle-aged subjects with acquired hearing loss. Br J Audiol 25:323-330, 1991.
- High WS, Fairbanks G, Glorig A. Scale for self assessment of hearing handicap. J Speech Hear Disord 29:215-230, 1964.
- Hu L, Bentler PM, Kano Y. Can test statistics in covariance structure analysis be trusted? Psychol Bull 112:351-362, 1992.
- Kalikow DN, Stevens KN, Elliot LL. Development of a test of speech intelligibility in Noise using sentence materials with controlled word predictability. J Acoust Soc Am 61:1337-1351, 1977.
- Karlsson AK, Rosenhall U. Aural rehabilitation in the elderly: Supply of hearing aids related to measured need and self assessed hearing problems. Scand Audiol 27:153-160, 1998.
- Kramer SE, Kapteyn TS, Festen JM, Tobi H. Factors in subjective hearing disability. Audiology 34:311-320, 1995.
- Lorenzi C, Gatehouse S, Lever C. Sound localization in normal-hearing listeners. J Acoust Soc Am 105:1-11, 1999a.
- Lorenzi C, Gatehouse S, Lever C. Sound localization in hearing-impaired listeners. J Acoust Soc Am 105:3454-3463, 1999b.
- Lormore KA, Stephens SDG. Use of the open-ended questionnaire with patients and their significant others. Br J Audiol 28:81-89, 1994.
- Long JS. Confirmatory Factor Analysis: A Preface to LISREL (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 33). Beverly Hills, CA: Sage, 1983.
- Lutman ME, Brown EJ, Coles RRA. Self-reported disability and handicap in the population in relation to pure-tone threshold, age, sex and type of hearing loss. Br J Audiol 21:45-58, 1987.

- MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling. *Psychol Meth* 1:130-149, 1996.
- MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. *Psychol Meth* 4:84-99, 1999.
- Marcus-Bernstein C. Audiologic and nonaudiologic correlates of hearing handicap in black elderly. *J Speech Hear Res* 29:301-312, 1986.
- Newman CW, Weinstein BE. Judgments of perceived hearing handicap by hearing-impaired elderly men and their spouses. *J Acad Rehab Audiol* 19:109-115, 1986.
- Newman CW, Weinstein BE, Jacobson GP, Hug GA. The hearing handicap inventory for adults: Psychometric adequacy and audiometric correlates. *Ear Hear* 11:430-433, 1990.
- Nilsson M, Soli SD, Sullivan JA. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am* 95:1085-1099, 1994.
- Noble WG, Atherley GRC. The Hearing Measure Scale: A questionnaire for the assessment of auditory disability. *J Aud Res* 10:229-250, 1970.
- Noble W, Byrne D, Ter Horst K. Auditory localization, detection of spatial separateness, and speech hearing in noise by hearing-impaired listeners. *J Acoust Soc Am* 102:2343-2352, 1997.
- Noble W, Ter-Horst K, Byrne D. Disabilities and handicaps associated with impaired auditory localization. *J Am Acad Audiol* 6:129-140, 1995.
- Nunnally JC. *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill, Inc., 1978.
- Nunnally JC, Bernstein IH. *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill, Inc., 1994.
- O'Mahoney CF, Stephens SDG, Cadge BA. Who prompts patients to consult about hearing loss? *Br J Audiol* 30:153-158, 1996.
- Pearsons KS, Bennett RL, Fidell S. Speech levels in various noise environments (EPA Report No. 600/1-77-025). Washington, DC: United States Environmental Protection Agency, 1977.
- Plomp R. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am* 63:533-549, 1978.
- Ringdahl A, Eriksson-Mangold M, Anderson G. Psychometric evaluation of the Gothenburg profile for measurement of experienced hearing disability and handicap: Applications with new hearing aid candidates and experienced hearing aid users. *Br J Audiol* 32:375-385, 1998.
- Rosenthal R, Rosnow RL. *Essentials of Behavioral Research: Methods and Data Analysis* (2nd ed.). New York: McGraw-Hill, 1991.
- Schow RL, Nerbonne MA. Assessment of hearing handicap by nursing home residents and staff. *J Acad Rehabil Audiol* 10:2-12, 1977.
- Schow RL, Nerbonne MA. Hearing Handicap and Denver scales: Applications, categories, interpretation. *J Acad Rehabil Audiol* 13:66-77, 1980.
- Schow RL, Nerbonne MA. Communication screening profile: Use with elderly clients. *Ear Hear* 3:135-147, 1982.
- Speaks C, Jerger J, Trammell J. Measurement of hearing handicap. *J Speech Hear Res* 13:768-776, 1970.
- Stephens D, France L, Lormore K. Effects of hearing impairment on the patient's family and friends. *Acta Otolaryngol (Stockh)* 115:165-167, 1995.
- Studebaker GA. A "rationalized" arcsine transform. *J Speech Hear Res* 28:455-462, 1985.
- Swan IRC, Gatehouse S. Factors influencing consultation for management of hearing disability. *Br J Audiol* 24:155-160, 1990.
- Walden BE, Demorest ME, Hepler EL. Self-report approach to assessing benefit derived from amplification. *J Speech Hear Res* 27:49-56, 1984.
- Widaman KF. Multitrait-multimethod models in aging research. *Exp Aging Res* 18:185-201, 1992.
- Wightman FL, Kistler DJ. Factors affecting the relative salience or sound localization cues. In: Gilkey RH, Anderson TR (eds): *Binaural and Spatial Hearing in Real and Virtual Environments*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 1-24, 1997.
- Woodcock RW. *Woodcock Reading Mastery Tests Manual*. Circle Pines, MN: American Guidance Service, Inc., 1973.